TACT: A Transfer Actor-Critic Learning Framework for Energy Saving in Cellular Radio Access Networks

Rongpeng Li*†, Zhifeng Zhao*†, Xianfu Chen‡, Jacques Palicot§ and Honggang Zhang*†

*York-Zhejiang Lab for Cognitive Radio and Green Communications

†Department of Information Science and Electronic Engineering

Zhejiang University, Zheda Road 38, Hangzhou 310027, China

Email: {lirongpeng, zhaozf, honggangzhang}@zju.edu.cn

‡VTT Technical Research Centre of Finland, P.O. Box 1100, FI-90571 Oulu, Finland

Email: xianfu.chen@vtt.fi

§Supélec, Avenue de la Boulaie CS 47601, Cesson-Sévigné Cedex, France

Email: jacques.palicot@supelec.fr

Abstract

Recent works have validated the possibility of energy efficiency improvement in radio access networks (RAN), depending on dynamically turn on/off some base stations (BSs). In this paper, we extend the research over BS switching operation, which should match up with traffic load variations. However, instead of depending on the predicted traffic loads, which is still quite challenging to precisely forecast, we firstly formulate the traffic variation as a Markov decision process (MDP). Afterwards, in order to foresightedly minimize the energy consumption of RAN, we adopt the actor-critic algorithm and design a reinforcement learning framework based BS switching operation scheme. Furthermore, to avoid the underlying curse of dimensionality in reinforcement learning, we propose a transfer actor-critic algorithm (TACT), which utilizes the transferred learning expertise in neighboring regions or

historical periods. The proposed TACT algorithm provably converges and contributes to a performance jumpstart. In the end, we evaluate our proposed scheme by extensive simulations under various practical configurations and prove the feasibility of significant energy efficiency improvement.

Index Terms

radio access networks, base stations, sleeping mode, green communications, energy saving, reinforcement learning, transfer learning, actor-critic algorithm

I. INTRODUCTION

The explosive popularity of smartphones and tablets has ignited a surging traffic load demand for radio access and has been incurring massive energy consumption and huge greenhouse gas (GHG) emission [1][2]. Specifically speaking, the information and communication technologies (ICT) industry accounts for 2% to 10% of the world's overall power consumption [3] and has emerged as one of the major contributors to the world-wide CO_2 emission. Besides that, there also exists economical pressure for cellular network operators to reduce the power consumption of their networks. It's envisioned that the power bill will doubly enlarge in fives years for China Mobile [4]. Meanwhile, the energy expenditure accounts for a significant proportion of the overall cost. Therefore, it's quite essential to improve the energy efficiency of ICT industry.

Currently, over 80% of the power consumption takes place in the radio access networks (RAN), especially the base stations (BSs) [5]. The reason behind this is largely due to that the present BS deployment is on the basis of peak traffic loads and generally stays active irrespective of the traffic load [6] while the traffic loads virtually vary heavily [7]. Recently, there has been a substantial body of work towards traffic load-aware BSs adaptation [8] and the authors have validated the possibility of energy efficiency improvement from different perspectives. Luca Chiaraviglio et al. [9] showed the possibility of energy saving by simulations. [10] and [11] proposed how to dynamically adjust the working status of BS, depending on the predicted traffic loads. However, to reliably predict the traffic loads is still quite challenging, which makes these works suffering in practical configurations. On the other hand, [12] and [13] presented dynamic BS switching algorithms with the traffic loads a prior and preliminarily proved the effectiveness of energy saving.

Besides, it is also found that turning on/off some of the BSs will immediately affect the BS,

with which a mobile terminal (MT) should be associated. Moreover, subsequent user's association choice in turn leads to the traffic load differences of BSs. Hence, any two consecutive BS switching operations are correlated with each other and current BS switching operation will also further influence the overall energy consumption in the long run. In other words, the expected energy saving scheme must be *foresighted* while minimizing the energy consumption. It should concern its effect on both the current and future system performance to deliver a visionary BS switching operation solution.

[6] presented a partially foresighted energy saving scheme which combines BS switching operation and user association by giving a heuristic solution on the basis of a stationary traffic load profile. In this paper, we try to solve these problem from a different perspective. Instead of predicting the volume of traffic loads, we apply Markov decision process (MDP) to model the traffic load variation. Afterwards, the solution to the formulated MDP model, i.e., BS switching operation (and corresponding user association as well) strategy, can be attained by making use of actor-critic algorithm [14][15], a reinforcement learning (RL) approach [16], one advantage of which is that there is no necessity to possess a prior knowledge about the traffic loads within the BSs. Within the reinforcement learning framework, a BS switching operation controller[1], as illustrated in Fig. 1, firstly estimates the traffic loads variation based on the on-line experience. Consequently, the controller can select one of the possible BS switching operations under the estimated circumstance and then decreases or increases the probability of the same action to be selected lately based on the needed cost. Here, the cost refers to the energy consumption due to such a BS switching operation. After repeating the actions and getting the corresponding cost, the controller would know how to choose the active BSs under one specific traffic load circumstance. Moreover, with the MDP model the resulting BS switching strategy is foresighted, which would improve energy efficiency in the long run.

However, some question may arise as the RL approaches usually suffer from *the curse of dimensionality* and master tasks with a large set of states and actions slowly [17][18]. Hence, a direct application of the RL approaches may sometimes get into trouble, because a BS switching operation controller usually takes charge of tens or even hundreds of BSs [11]. In this paper,

[1]In practice, such a centralized BS switching operation can be conducted by the base station controller (BSC) in second generation (2G) cellular networks or the radio network controller (RNC) in third generation (3G) or long term evolution (LTE) cellular networks. In this paper, we generalize it as a BS switching operation controller.

we deal with the application problem by utilizing the conceptual idea of transfer learning (TL) [19]-[22]. TL, which mostly concern how to recognize and apply the knowledge learned from one or more previous tasks (*source tasks*) to more effectively learn to solve novel task (*target task*) [20], is intuitively appealing, cognitive inspired, and has led to a burst of research activities. Meanwhile, the spatial and temporal relevancy in the traffic loads [23] make it meaningful to transfer the learned BS switching operation strategy in neighboring region at historical moments (source task) to help speed up the learning process in regions of interest (target task) as depicted in Fig 1. As a result, the learning framework of BS switching operation is further addressed by incorporating the idea of TL into the classical actor-critic algorithm (AC) and present a Transfer Actor-CriTic algorithm (TACT) in this paper.

In a nutshell, our work proposes a reinforcement learning framework to energy saving scheme in RANs. Beyond that, compared to the previous work, this paper provides the following three key insights:

- Firstly, we show that the learning framework scheme is feasible to save the energy consumption in RANs without the knowledge of traffic loads a prior. Moreover, the performance of the learning framework scheme approaches that of the state of the art scheme (SOTA), which is assumed to have fully knowledge of traffic loads. These preliminary results have already been presented in [24].
- Secondly, we extend the idea of TL to the conventional RL algorithms and show that the proposed transfer actor-critic algorithm (TACT) outperforms the classical AC algorithm with a performance jumpstart.
- Thirdly, this paper details the convergence analysis of the TACT algorithm and thereby contributes to the general literature in RL filed, especially the general AC algorithm.

The remainder of the paper is organized as follows. In Section II, we introduce the system model and formulate the traffic variation as an MDP. In Section III, we talk about energy saving scheme by the conventional RL framework. Section IV focuses on the incorporation of idea of TL into the conventional RL framework and investigates the convergence proof of the transfer actor critic algorithm. Section V evaluates the proposed schemes and presents the validity and effectiveness. Finally, we present a conclusion of this paper in Section VI .

II. System model and problem formulation

A. System model

An RAN usually consists of multiple BSs while the traffic loads of BSs are usually fluctuating, thus often making BSs under-utilization. In this paper, let's assume that there exists a region $\mathcal{L} \in \mathbb{R}^2$ served by a set of overlapped BSs $\mathcal{B} = \{1, \ldots, N\}$ as Fig. 1 depicts, i.e., Region 1 or Region 2. In addition, there exists a BS switching operation controller, which can timely know the traffic loads in these BSs at current stage and correspondingly determine the energy efficient working status of any BS (i.e., active/sleeping mode) at next stage in a centralized way. Beyond that, the paper focuses on downlink communication, i.e., from BSs to MTs. Meanwhile, the file transmission requests at a location $x \in \mathcal{L}$ arrive following a Poisson point process with arrival rate per unit area $\lambda(x)$ and file size $\frac{1}{\mu(x)}$. After that, the *traffic load density* at a location $x \in \mathcal{L}$ is defined as $\gamma(x) = \lambda(x)/\mu(x) < \infty$ [6]. Therefore, the traffic load density can capture the spatial traffic variations. For example, a hot spot can be characterized by a high arrival rate and/or possibly large file sizes. Hence, when the set of BSs $\mathcal{B}_{on}$ is turned on, the traffic loads severed by BS $i \in \mathcal{B}_{on}$ can be represented as $\Gamma_i = \int_{\mathcal{L}} \gamma(x) I_i(x, \mathcal{B}_{on})\, \mathrm{d}x$, whereas $I_i(x, \mathcal{B}_{on}) = 1$ is a user association indicator and denotes location x is served by BS $i \in \mathcal{B}_{on}$ and vice versa. Otherwise, if a BS i is in sleeping mode, i.e., $i \in \mathcal{B} \setminus \mathcal{B}_{on}$, the traffic load is defined as zero, namely $\Gamma_i = 0$. To demonstrate the traffic load variation condition, i.e., $p(\Gamma_i^{k+1} | \Gamma_i^k)$, we use a finite state Markov process (FSMC). Moreover, the traffic load Γ_i for BS i is partitioned into two parts by a boundary point Γ_b. Here, Γ_b can be the average traffic loads in one BS over a certain period, thus feasible to be known in advance based on the historical records. Therefore, the traffic loads for a specific BS have merely two states, i.e., $s_i = 0$ if $\Gamma_i < \Gamma_b$ and $s_i = 1$ if $\Gamma_i > \Gamma_b$. Subsequently, a state vector $\mathbf{s} = \{s_1, \cdots, s_N\} \in \mathbb{S} = S_1 \times \cdots \times S_N$ is constructed to model the traffic load variation for the region of interest.

Let's denote the transmission rate of a user located at x and served by BS $i \in \mathcal{B}_{on}$ as $c_i(x, \mathcal{B}_{on})$. For analytical convenience, assume that $c_i(x, \mathcal{B}_{on})$ does not change over time, i.e., we do not consider fast fading or dynamic inter-cell interferences. Instead, $c_i(x, \mathcal{B}_{on})$ is assumed as a time-averaged transmission rate in this paper, based on the fact that the time scale of user association is commonly much larger than the time scale of fast fading or dynamic inter-cell interferences. Hence, the inter-cell interference is considered as static Gaussian-like noise, which is feasible

under interference randomization or fractional frequency reuse, also consistent with the model in [6][25]. Beyond that, though $c_i(x, \mathcal{B}_{on})$ is location-dependent, it is not necessarily determined by the distance from the BS i due to the shadowing effect.

Furthermore, the *system load density* can be defined as the fraction of time required to deliver traffic load $\gamma(x)$ from BS $i \in \mathcal{B}_{on}$ to location x, namely $\varrho_i(x) = \gamma(x)/c_i(x, \mathcal{B}_{on})$. Analogous to the definition of traffic load, the system load for an active BS $i \in \mathcal{B}_{on}$ can be represented as $\rho_i = \int_{\mathcal{L}} \varrho_i(x) I_i(x, \mathcal{B}_{on})\, \mathrm{d}x$. Meanwhile, the system load for a sleeping BS i is defined as zero, namely $\rho_i = 0,\ \text{if } i \in \mathcal{B} \setminus \mathcal{B}_{on}$. Hence, the indicator set $\mathbb{I} = \{I_i(x, \mathcal{B}_{on}) | i \in \mathcal{B}, x \in \mathcal{L}\}$ is feasible [26] if one BS can serve $\rho_i < 1, \forall i \in \mathcal{B}$. Eventually, our goal is to choose certain active BSs and find a feasible user association indicator set to minimize the overall energy consumption. By exploiting the proposed learning framework, the controller can know the BS switching operation strategy at last without the prior knowledge of traffic loads. We will give the details in Section III.

B. Problem formulation

In this paper, we primarily aim to minimize the whole-scale energy consumption of BSs in RANs. Our previous work [11] has shown the energy consumption of BS is not linearly proportional to the traffic load within its coverage area. Moreover, the energy consumption of BSs consists of two categories: constant one and variant one that is proportional to BS's traffic load. Hence, we adopt the generalized energy consumption model [6], which can be summarized as

$$\psi(\rho, \mathcal{B}_{on}) = \sum_{i \in \mathcal{B}_{on}} \left[(1 - q_i)\rho_i P_i + q_i P_i\right], \tag{1}$$

where $\rho = \{\rho_1, \cdots, \rho_N\}$. Besides, $q_i \in (0, 1)$ is the portion of constant power consumption for BS i, and P_i is the maximum power consumption of BS i when it is fully utilized.

Above all, our problem is to find an optimal set of active BSs and corresponding user association that minimizes the function of the energy consumption, namely

$$\begin{aligned} &\min_{\mathcal{B}_{on}, \rho} \left\{\psi(\rho, \mathcal{B}_{on})\right\}, \\ &s.t.\ \rho_i \in [0, 1)\ \forall i \in \mathcal{B}. \end{aligned} \tag{2}$$

III. Stochastic BS switching operation in reinforcement learning framework

A. Markov decision process

An MDP is defined as a tuple $M =< \mathbb{S}, \mathbb{A}, p, C >$, where $\mathbb{S}$ is the state space, $\mathbb{A}$ is the action space, p is a state transition probability function, and C is a cost function[2]. Specifically, at stage k, the traffic load state is $\mathbf{s}^k$. The controller choose to turn some BSs into sleeping mode (Action $\mathbf{a}^k$) and the users correspondingly associate themselves with the remaining active BSs according to an indicator set $\mathbb{I}^k$[3]. Thereafter, the traffic load state will transform into $\mathbf{s}^{k+1}$ with probability $p(\mathbf{s}^{k+1}|\mathbf{s}^k, \mathbf{a}^k)$. Meanwhile, the immediate cost generated by the environment (computed by Equation (1)) is fed back to the agent, i.e., the BS switching operation controller.

The goal is to find a strategy π, which maps a state $\mathbf{s}$ to an action $\pi(\mathbf{s})$, i.e., $\mathbf{a}^k$, to minimize the discounted accumulative cost starting from the state $\mathbf{s}$. Formally, this accumulative cost is called as a state value function, which can be calculated by [16]

$$\begin{aligned} V^{\pi}(\mathbf{s}) &= E_{\pi}\left[\sum_{k=0}^{\infty}\gamma^k C(\mathbf{s}^k, \pi(\mathbf{s}^k)|\mathbf{s}^0 = \mathbf{s})\right] \\ &= E_{\pi}\left[C(\mathbf{s}, \pi(\mathbf{s})) + \gamma\sum_{\mathbf{s}'\in\mathbb{S}} p(\mathbf{s}'|\mathbf{s}, \pi(\mathbf{s}))V^{\pi}(\mathbf{s}')\right], \end{aligned} \tag{4}$$

where γ is the discount factor that maps the future cost to the current state. Given the diminishing importance of future cost than the current one, γ is smaller than 1. The optimal strategy π^*

[2]It may be a reward function R on the basis of specific research scenarios. Moreover, it's worthwhile to note here that we use the lowercased $c_i(x, \mathcal{B}_{on})$ to denote transmission rate from BS i to location x while the uppercased C denotes the cost function.

[3]In this paper, we adopt and modify the approach for user association in [6]. At stage k, the user association set $\mathbb{I}^k$, which achieves the minimization of total cost, would be that users at location x choose to join BS i^*, while i^* satisfies

$$i^*(x) = \arg\max_{j\in\mathcal{B}_{on}} \frac{c_j(x, \mathcal{B}_{on})}{(1-q_j)P_j},\ \forall x \in \mathcal{L}. \tag{3}$$

Intuitively, Equation (3) means that users at location x prefer to choose to join the BS with the largest transmission rate at the same traffic load-variant power consumption.

It's worthwhile to note here that this user association scheme may degrade the quality of experience (QoE), such as increasing the delay, etc. We leave how to strike the balance between the user QoE and energy consumption as future work.

satisfies the Bellman equation [16]:

$$\begin{aligned} V^*(\mathbf{s}) &= V^{\pi^*}(\mathbf{s}) \\ &= \min_{\mathbf{a}\in\mathbb{A}} \left\{ E_{\pi^*} \left[C(\mathbf{s},\mathbf{a}) + \gamma \sum_{\mathbf{s}'\in\mathbb{S}} p(\mathbf{s}'|\mathbf{s},\mathbf{a}) V^{\pi^*}(\mathbf{s}') \right] \right\}. \end{aligned} \tag{5}$$

Since the optimal strategy not only minimizes the current cost, but the cumulative cost from the beginning, it contributes to design a foresighted energy saving scheme.

B. The actor-critic learning framework for energy saving scheme

There have been some well-known methods to solve the MDP issues such as dynamic programming [16]. Unfortunately, these methods heavily depends on prior knowledge of the environmental dynamics. However, it's challenging to know the future traffic loads precisely in advance. Therefore, in this paper, we employ an actor-critic algorithm, one kind of reinforcement learning to solve the MDP problem. The reasons to adopt actor-critic algorithm are twofold [27]: (i) since it generates the action directly from the stored policy, it requires little computation to select an action to perform; (ii) it can learn an explicitly stochastic policy which may be useful in non-Markov traffic variation environment of RAN.

As the name suggests, the actor-critic algorithm encompasses three components: actor, critic, and environment as illustrated in Fig. 2 (Left). At a given state, the actor selects an action in a stochastic way and then executes it. This execution transforms the state of environment to a new one with a certain probability, and feeds back the cost to the actor. Then, the critic criticizes the action executed by the actor through a time difference (TD) error. After the criticism, the actor will prefer to select the action yielding a smaller cost with a higher tendency, and vice versa. The algorithm repeats the above procedure until convergence.

We design an actor-critic learning framework for energy saving scheme as illustrated in Fig. 3.

1) Action selection: Beforehand, let's assume that when the controller needs to select an action, the system is at the beginning of stage k. Meanwhile, the traffic load state is $\mathbf{s}^k$. Thereafter, the controller selects an action according to a stochastic strategy, the purpose of which is to improve performance while explicitly balancing two competing objectives: a) searching for a better BS switching operation (exploration) and b) taking as little cost as possible (exploitation), such that the controller not only performs the good BS switching operation based on its past experience

but also is able to explore a new one. The most common methodology is to use a Boltzmann distribution. The controller chooses an action $\mathbf{a}$ in state $\mathbf{s}^k$ of stage k with probability [16]

$$\pi^k(\mathbf{s}^k, \mathbf{a}) = \frac{\exp\{p(\mathbf{s}^k, \mathbf{a})/\tau\}}{\sum_{\mathbf{a}' \in \mathbb{A}} \exp\{p(\mathbf{s}^k, \mathbf{a}')/\tau\}}, \tag{6}$$

where τ is a positive parameter called the temperature. In addition, $p(\mathbf{s}^k, \mathbf{a}^k)$ indicates the tendency to select action $\mathbf{a}^k$ at the state $\mathbf{s}^k$, and it will update itself after every iteration. It's worthwhile to note that though there exists the possibility that the remaining active BSs are not enough to serve the traffic loads in the present stage k, the controller can start an emergent response paradigm to quickly turn on some BSs in this case as the conventional energy saving scheme commonly does, which is out of the scope of this paper. Hence, in this paper, we assume the action $\mathbf{a}^k$, which the controller finally chooses, can meet the traffic load requirement.

(2) User association and data transmission: After the controller chooses to turn some of BSs into sleeping mode, the users at location x choose to connect one BS according to Equation (3) and start the data communication.

(3) State-value function update: After the transmission part of stage k, the traffic loads in each BS will change, thus transforming the system to state $\mathbf{s}^{k+1}$. Meanwhile, the total cost for the transmission would be $C^k(\mathbf{s}^k, \mathbf{a}^k)$. Consequently, a TD error $\delta(\mathbf{s}^k, \mathbf{a}^k)$ would be computed by the difference between the state-value function $V^k(\mathbf{s}^k)$ estimated at the preceding state and the one $C^k(\mathbf{s}^k, \mathbf{a}^k) + \gamma \cdot V^k(\mathbf{s}^{k+1})$ at the critic, namely

$$\begin{aligned} \delta^k(\mathbf{s}^k, \mathbf{a}^k) &= C^k(\mathbf{s}^k, \mathbf{a}^k) + \gamma \sum\nolimits_{\mathbf{s}' \in \mathbb{S}} p(\mathbf{s}'|\mathbf{s}^k, \mathbf{a}^k) V(\mathbf{s}') - V(\mathbf{s}^k) \\ &= C^k(\mathbf{s}^k, \mathbf{a}^k) + \gamma \cdot V^k(\mathbf{s}^{k+1}) - V^k(\mathbf{s}^k). \end{aligned} \tag{7}$$

Afterwards, the TD error would feed back to the actor. By the way, the state-value function would be updated as

$$V^{k+1}(\mathbf{s}^k) = V^k(\mathbf{s}^k) + \alpha(\nu_1(\mathbf{s}^k, k)) \cdot \delta^k(\mathbf{s}^k, \mathbf{a}^k). \tag{8}$$

Here, $\nu_1(\mathbf{s}^k, k)$ denotes the occurrence times of state $\mathbf{s}^k$ in these k stages. $\alpha(n)$ is a positive step-size parameter that affects the convergence rate. On the other hand, if $\mathbf{s} \neq \mathbf{s}^k$, $V^{k+1}(\mathbf{s})$ will be kept the same as $V^k(\mathbf{s})$, namely $V^{k+1}(\mathbf{s}) = V^k(\mathbf{s}), \forall \mathbf{s} \in \mathbb{S}$ but $\mathbf{s} \neq \mathbf{s}^k$.

(4) Policy update: At the end of stage k, the critic would employ the TD error to "criticize" the selected action, which is implemented as

$$p^{k+1}(\mathbf{s}^k, \mathbf{a}^k) = p^k(\mathbf{s}^k, \mathbf{a}^k) - \beta(\nu_2(\mathbf{s}^k, \mathbf{a}^k, k)) \cdot \delta^k(\mathbf{s}^k, \mathbf{a}^k), \tag{9}$$

Similar to $\nu_1(\mathbf{s}^k, k)$, $\nu_2(\mathbf{s}^k, \mathbf{a}^k, k)$ indicates the executed times of action $\mathbf{a}^k$ at state $\mathbf{s}^k$ in these k stages. $\beta(n)$ is a positive step-size parameter. Equation (6) and Equation (9) ensure one action under a specific state can be selected with higher probability if the "foresighted" cost it takes is comparatively smaller, i.e., $\delta(\mathbf{s}^k) < 0$. Additionally, if $\mathbf{a} \neq \mathbf{a}^k$, $p^{k+1}(\mathbf{s}^k, \mathbf{a})$ will remain unchanged, namely $p^{k+1}(\mathbf{s}^k, \mathbf{a}) = p^k(\mathbf{s}^k, \mathbf{a}), \forall \mathbf{a} \in \mathbb{A}$ but $\mathbf{a} \neq \mathbf{a}^k$.

If each action is executed infinitely often in every state, in other words, if in the limit, the learning strategy is greedy with infinite exploration, the value function $V(\mathbf{s})$ and strategy $\pi^k(\mathbf{s}, \mathbf{a})$ will finally converge to V^* and π^* with probability (w.p.) 1 as $k \to \infty$ [28].

IV. Transfer actor-critic algorithm for stochastic BS switching operation

A. Motivation and formulation of transfer actor-critic algorithm

The previous section addresses the methodology to exploit the classical AC algorithm to conduct the BS switching operation, culminating in an effective energy saving strategy in the end. In this section, we present the means that the controller utilizes the knowledge of learned strategy in a neighboring region or a historical period to help itself be in the groove of finding the optimal BS switching operation.

Basically, the policy, say $p(\mathbf{s}, \mathbf{a})$, which finally determines the strategy $\pi(\mathbf{s}, \mathbf{a})$ in one learning task, indicates the tendency of action $\mathbf{a}$ to be chosen in state $\mathbf{s}$. When the learning process converges, the tendency to choose a specific action $\mathbf{a}$ is comparatively larger than that of other actions. In other words, it means that if the controller decides the BS switching operation according to action $\mathbf{a}$, the energy consumption reduction in the whole system is tending to be optimized in the long run. Hence, if the knowledge of this policy $p(\mathbf{s}, \mathbf{a})$ is transferred to another task, i.e., the knowledge transferred from Region 1 (source task) to Region 2 (target task) in Fig. 1, the controller in the target task can make an attempt by taking the same action $\mathbf{a}$ when the traffic loads come into state $\mathbf{s}$. Compared to learning from the scratch, the controller might directly make the wisest choice at the very beginning. However, in spite of the similarities between the source task and the target task, there still exist the differences. For example, the system might come into the same state in two different tasks, whereas the traffic loads in the source task (i.e., Region 1) might be usually higher than that in the target one (i.e, Region 2). Hence, instead of staying on the chosen action $\mathbf{a}$, the controller can make a more aggressive choice to turn more BSs into sleeping mode, thus saving more energy consumption. Consequently,

in this case, the transferred policy guides in a negative manner. To avoid this underlying problem, the transferred tendency should have a decreasing impact on choosing a certain action once the controller has attempted to choose this action and nurtured its own learning experience.

Afterwards, we propose a new policy update method for Transferred Actor-CriTic algorithm (TACT) as Fig. 2. In the TACT algorithm, the overall policy to select an action p_o is divided as the native one p_n and the exotic one p_e. Without loss of generality, let's assume that at stage k, the traffic load state is $\mathbf{s}^k$ and the chosen action is $\mathbf{a}^k$. Accordingly, the overall policy p_o is updated as

$$p_o^{k+1}(\mathbf{s}^k, \mathbf{a}^k) = \left[(1 - \zeta(\nu_2(\mathbf{s}^k, \mathbf{a}^k, k)))p_n^{k+1}(\mathbf{s}^k, \mathbf{a}^k) + \zeta(\nu_2(\mathbf{s}^k, \mathbf{a}^k, k))p_e(\mathbf{s}^k, \mathbf{a}^k)\right]_{-p_t}^{p_t}, \tag{10}$$

where $[x]_a^b$ with $b > a$, denotes the Euclidean projection of x onto the interval [a,b], i.e., $[x]_a^b = a$ if $x < a$, $[x]_a^b = b$ if $x > b$, and $[x]_a^b = x$ if $a \leq x \leq b$. In this case, $a = -p_t$ and $b = p_t$, with $p_t > 0$. Additionally, if $\mathbf{a} \neq \mathbf{a}^k$, $p_o^{k+1}(\mathbf{s}^k, \mathbf{a})$ will remain unchanged, namely $p_o^{k+1}(\mathbf{s}^k, \mathbf{a}) = p_o^k(\mathbf{s}^k, \mathbf{a}), \forall \mathbf{a} \in \mathbb{A}$ but $\mathbf{a} \neq \mathbf{a}^k$. Besides that, $p_n(\mathbf{s}, \mathbf{a})$ still updates itself according to the classical actor-critic algorithm, namely Equation (9).

Initially, the exotic policy $p_e(\mathbf{s}, \mathbf{a})$ dominates in the overall strategy. Hence, when the environment enters a state $\mathbf{s}$, the presence of $p_e(\mathbf{s}, \mathbf{a})$ contributes to choose the action, which might be optimal to $\mathbf{s}$ in the source task. Consequently, the proposed tendency update method leads to a possible performance jumpstart. Beyond that, $\zeta(n) \in (0, 1)$ is the transfer rate and $\zeta(n) \to 0$ as $n \to \infty$. The existence of $\zeta(n)$ continuously decreases the effect of the transferred exotic policy $p_n(\mathbf{s}, \mathbf{a})$. Therefore, the controller can not only take advantage of the learned expertise in the source task, but also swiftly get rid of the negative guidelines.

Finally, we summarize our proposed TACT algorithm in Algorithm 1 .

B. Convergence analysis

Next, we are interested in the convergence of TACT algorithm. We start the analysis by introducing several related lemmas. Singh [28] shows that the Boltzmann method is greedy in the limit with infinite exploration, based on a large enough τ. Therefore, we have the following lemma.

Lemma 1. If we use the Boltzmann exploration method with a large enough τ, there thereby

Algorithm 1 TACT : The Transfer Learning Framework for Energy Saving Scheme

Initialization:

for each $\mathbf{s} \in \mathbb{S}$, each $\mathbf{a} \in \mathbb{A}$ **do**

Initialize state-value function $V(\mathbf{s})$, native policy function $p_n(\mathbf{s}, \mathbf{a})$, and strategy function $\pi(\mathbf{s}, \mathbf{a})$;

end for

Repeat until convergent

1) Choose an action $\mathbf{a}^k$ in state $\mathbf{s}^k$ according to $\pi(\mathbf{s}^k, \mathbf{a}^k)$;
2) Users at location x connect one BS i by $i^*(x) = \arg\max_{j \in \mathcal{B}_{on}} \frac{c_j(x, \mathcal{B}_{on})}{(1-q_j)P_j}$, $\forall x \in \mathcal{L}$ and then start data transmission;
3) If $\rho_i \leq 1, \forall i \in \mathcal{L}$, the chosen action is feasible. The cost function $C(\mathbf{s}^k, \mathbf{a}^k)$ is calculated by $\sum_{i \in \mathcal{B}_{on}} [(1-q_i)\rho_i P_i + q_i P_i]$; otherwise, an emergent response paradigm starts as the conventional scheme does.
4) Identify the traffic loads and accordingly update state $\mathbf{s}^k \to \mathbf{s}^{k+1}$ and compute the TD error by $\delta^k(\mathbf{s}^k) = C^k(\mathbf{s}^k, \mathbf{a}^k) + \gamma \cdot V^k(\mathbf{s}^{k+1}) - V^k(\mathbf{s}^k)$;
5) Update the state-value function $V(\mathbf{s}^k)$ by $V^{k+1}(\mathbf{s}^k) = V^k(\mathbf{s}^k) + \alpha(\nu_1(\mathbf{s}^k, k)) \cdot \delta^k(\mathbf{s}^k)$;
6) Update the native tendency function $p_n(\mathbf{s}^k, \mathbf{a}^k)$ by $p_n^{k+1}(\mathbf{s}^k, \mathbf{a}^k) = p_n^k(\mathbf{s}^k, \mathbf{a}^k) - \beta(\nu_2(\mathbf{s}^k, \mathbf{a}^k, k)) \cdot \delta^k(\mathbf{s}^k, \mathbf{a}^k)$, and update the function $p_o(\mathbf{s}^k, \mathbf{a}^k)$ by $p_o^{k+1}(\mathbf{s}^k, \mathbf{a}^k) = \left[(1 - \zeta(\nu_2(\mathbf{s}^k, \mathbf{a}^k, k)))p_n^{k+1}(\mathbf{s}^k, \mathbf{a}^k) + \zeta(\nu_2(\mathbf{s}^k, \mathbf{a}^k, k))p_e(\mathbf{s}^k, \mathbf{a}^k)\right]_{-p_t}^{p_t}$;
7) Update the strategy function $\pi^{k+1}(\mathbf{s}^k, \mathbf{a}) = \frac{\exp\{p_o^{k+1}(\mathbf{s}^k, \mathbf{a})/\tau\}}{\sum_{\mathbf{a}' \in \mathbb{A}} \exp\{p_o^{k+1}(\mathbf{s}^k, \mathbf{a}')/\tau\}}$, for all $\mathbf{a} \in \mathbb{A}$.

exists an $\eta > 0$, such that

$$\lim_{k \to \infty} \frac{\nu_2(\mathbf{s}, \mathbf{a}, k)}{k} \geq \eta, \forall \mathbf{s} \in \mathbb{S}, \mathbf{a} \in \mathbb{A}. \tag{11}$$

In other words, as $k \to \infty$, $\nu_2(\mathbf{s}, \mathbf{a}, k) = \eta k \to \infty$.

Definition 1. Define a function $\vartheta_{\mathbf{s},\mathbf{a}}(p_o)$ as

$$\vartheta_{\mathbf{s},\mathbf{a}}(p_o) = \begin{cases} 0 \text{ if } p_o(\mathbf{s}, \mathbf{a}) = p_t & \text{and } \delta(\mathbf{s}, \mathbf{a}) \geq 0, \\ \quad \text{or } p_o(\mathbf{s}, \mathbf{a}) = -p_t & \text{and } \delta(\mathbf{s}, \mathbf{a}) \leq 0, \\ 1 \text{ otherwise.} \end{cases} \tag{12}$$

The next theorem states that our proposed policy update tracks an ordinary differential equation (ODE).

Theorem 1. Assume that the learning rate $\beta(n)$ in Equation (9) satisfies

$$\sum_{n=0}^{\infty}\beta(n)=\infty,\beta(n)\geq 0,\sum_{n=0}^{\infty}\beta(n)^2<\infty, \tag{13}$$

and the transfer rate $\zeta(n)$ satisfies $\lim\zeta(n)/\beta(n)\to 0$ as $n\to\infty$. $p_o(\mathbf{s},\mathbf{a})$ asymptotically track the solution of the ODE

$$\dot{p}_o(\mathbf{s},\mathbf{a})(t)=-\delta(\mathbf{s},\mathbf{a})\vartheta_{\mathbf{s},\mathbf{a}}(p_o),\forall\mathbf{s}\in\mathbb{S},\mathbf{a}\in\mathbb{A}, \tag{14}$$

where $\delta(\mathbf{s},\mathbf{a})=\lim\delta^k(\mathbf{s},\mathbf{a})$ as $k\to\infty$.

Proof: The proof is given in Appendix. ■

In addition, we introduce the definition of a *strict Lyapunov function* [29], which is the fundamental of our following proof.

Definition 2. Suppose that for an ODE $\dot{z}(t)=f(z)$ defined on a region $\mathcal{D}$, $V(z)$ is a continuously differentiable and real-valued function of z such that $V(0)=0,V(z)>0,\forall z\neq 0$, if $\dot{V}(t)=\nabla V\cdot\dot{z}(t)=\nabla V\cdot f(z)\leq 0$ on the region $\mathcal{D}$, and the equality holds only when $\dot{z}(t)=0$, the function $V(z)$ is a strict Lyapunov function for the ODE $\dot{z}(t)$.

Our proof relies on the following theorem by Konda and Borkar [14], which establishes the convergence of a general actor-critic algorithm.

Theorem 2. Assume that the learning rate $\alpha(n)$ satisfies the assumptions in Section 2.2 [14] and $\beta(n)$ and $\zeta(n)$ meet the conditions in Theorem 1. If the strategy π, which is derived by Equation (6) with the policy update method given by Equation (10), has a *strict Lyapunov function* for the ODE $\dot{\pi}(t)$, we thereby have π converges w.p. 1 and $\|\pi-\pi^*\|\leq\epsilon$ for any $\epsilon>0$ as $p_t\to\infty$.

Beforehand, it comes the following lemma by directly applying Equation (4) in Equation (7).

Lemma 2.

$$\sum_{\mathbf{a}\in\mathbb{A}}\delta(\mathbf{s},\mathbf{a})\pi(\mathbf{s},\mathbf{a})=0,\forall\mathbf{s}\in\mathbb{S}. \tag{15}$$

Lemma 3. If the strategy $\pi(\mathbf{s},\mathbf{a})$ tracks the solution of ODE $\dot{\pi}(\mathbf{s},\mathbf{a})(t)$, and $\dot{\pi}(\mathbf{s},\mathbf{a})(t)$ satisfies $\dot{\pi}(\mathbf{s},\mathbf{a})(t)\delta(\mathbf{s},\mathbf{a})\leq 0$, then we have $\nabla V^{\pi}(\mathbf{s})\dot{\pi}(\mathbf{s},\mathbf{a})(t)\leq\dot{\pi}(\mathbf{s},\mathbf{a})(t)\delta(\mathbf{s},\mathbf{a})\leq 0,\forall\mathbf{s}\in\mathbb{S}$.

Proof: For two distinct policies π and π', let's define a value function operation $T(\pi', V^{\pi}(\mathbf{s}))$ $= E_{\pi'}\left[C(\mathbf{s},\mathbf{a}) + \gamma \sum_{\mathbf{s}'\in\mathbb{S}} p(\mathbf{s}'|\mathbf{s},\mathbf{a})V^{\pi}(\mathbf{s}')\right]$. Assume that there exists an infinitesimal $\epsilon > 0$ such that $\pi + \epsilon\dot{\pi}(t)$ is still a valid strategy. If denote $\pi' = \pi + \epsilon\dot{\pi}(t)$, we thereby have

$$
\begin{aligned}
T(\pi', V^{\pi}(\mathbf{s})) - V^{\pi}(\mathbf{s}) &= E_{\pi'}\left[C(\mathbf{s},\mathbf{a}) + \gamma \sum_{\mathbf{s}'\in\mathbb{S}} p(\mathbf{s}'|\mathbf{s},\mathbf{a})V^{\pi}(\mathbf{s}')\right] - V^{\pi}(\mathbf{s}) \\
&= \sum_{\mathbf{a}\in\mathbb{A}}\left\{(\pi(\mathbf{s},\mathbf{a}) + \epsilon\dot{\pi}(\mathbf{s},\mathbf{a}))\left[C(\mathbf{s},\mathbf{a}) + \gamma \sum_{\mathbf{s}'\in\mathbb{S}} p(\mathbf{s}'|\mathbf{s},\mathbf{a})V^{\pi}(\mathbf{s}') - V^{\pi}(\mathbf{s})\right]\right\} \\
&= \sum_{\mathbf{a}\in\mathbb{A}}(\pi(\mathbf{s},\mathbf{a}) + \epsilon\dot{\pi}(\mathbf{s},\mathbf{a}))\delta(\mathbf{s},\mathbf{a}) \\
&= \sum_{\mathbf{a}\in\mathbb{A}}\epsilon\dot{\pi}(\mathbf{s},\mathbf{a})\delta(\mathbf{s},\mathbf{a}) \leq 0
\end{aligned}
$$

The last equality follows from Lemma 2.

Denote an iteration operation of $T(\pi', V^{\pi}(\mathbf{s}))$ as $T^{n}(\pi', V^{\pi}(\mathbf{s})) = T^{n-1}(\pi', T(\pi', V^{\pi}(\mathbf{s})))$, we have $T^{n}(\pi', V^{\pi}(\mathbf{s})) \leq T^{n-1}(\pi', V^{\pi}(\mathbf{s})) \leq \cdots \leq V^{\pi}(\mathbf{s})$.

In addition, $T^{n}(\pi', V^{\pi}(\mathbf{s})) - V^{\pi}(\mathbf{s}) \leq \sum_{\mathbf{a}\in\mathbb{A}} \epsilon\dot{\pi}(\mathbf{s},\mathbf{a})\delta(\mathbf{s},\mathbf{a})$, for $n > 1$. As $n \to \infty$, $T(\pi', V^{\pi}(\mathbf{s}))$ $\to V^{\pi'}(\mathbf{s})$, we obtain

$$
\frac{V^{\pi'}(\mathbf{s}) - V^{\pi}(\mathbf{s})}{\epsilon} = \frac{V^{\pi+\epsilon\dot{\pi}}(\mathbf{s}) - V^{\pi}(\mathbf{s})}{\epsilon} \leq \dot{\pi}(\mathbf{s},\mathbf{a})\delta(\mathbf{s},\mathbf{a}) \leq 0.
$$

As $\epsilon \to 0$, $\nabla V^{\pi}(\mathbf{s})\dot{\pi}(\mathbf{s},\mathbf{a})(t) \leq \dot{\pi}(\mathbf{s},\mathbf{a})\delta(\mathbf{s},\mathbf{a}) \leq 0$. The claim follows. ■

Theorem 3. $\sum_{\mathbf{s}\in\mathbb{S}} V^{\pi}(\mathbf{s})$ is a strict Lyapunov function for ODE $\dot{\pi}(t)$.

Proof: By explicit differentiating Equation (6) over $p_o(\mathbf{s},\mathbf{a})(t)$, we have

$$
\begin{aligned}
\dot{\pi}(t) &= \frac{\frac{1}{\tau}\exp\left[p_o(\mathbf{s},\mathbf{a})/\tau\right]}{\sum_{\mathbf{a}'\in\mathbb{A}}\exp\left[p_o(\mathbf{s},\mathbf{a}')/\tau\right]}\dot{p}_o(\mathbf{s},\mathbf{a}) - \frac{\frac{1}{\tau}\exp\left[p_o(\mathbf{s},\mathbf{a})/\tau\right]\sum_{\mathbf{a}'\in\mathbb{A}}\left\{\exp\left[p_o(\mathbf{s},\mathbf{a}')/\tau\right]\dot{p}_o(\mathbf{s},\mathbf{a}')\right\}}{\left\{\sum_{\mathbf{a}'\in\mathbb{A}}\exp\left[p_o(\mathbf{s},\mathbf{a}')/\tau\right]\right\}^2} \\
&= \frac{1}{\tau}\pi(\mathbf{s},\mathbf{a})\dot{p}_o(\mathbf{s},\mathbf{a}) - \frac{1}{\tau}\pi(\mathbf{s},\mathbf{a})\sum_{\mathbf{a}'\in\mathbb{A}}\pi(\mathbf{s},\mathbf{a}')\dot{p}_o(\mathbf{s},\mathbf{a}') \\
&= \frac{1}{\tau}\pi(\mathbf{s},\mathbf{a})\dot{p}_o(\mathbf{s},\mathbf{a}) - \frac{1}{\tau}\pi(\mathbf{s},\mathbf{a})\sum_{\mathbf{a}'\in\mathbb{A}}\pi(\mathbf{s},\mathbf{a}')\left[-\delta(\mathbf{s},\mathbf{a}')\vartheta_{\mathbf{s},\mathbf{a}'}(p_o)\right] \\
&= \frac{1}{\tau}\pi(\mathbf{s},\mathbf{a})\dot{p}_o(\mathbf{s},\mathbf{a}).
\end{aligned}
$$

The last equality follows from Lemma 2. By Theorem 1,

$$
\dot{\pi}(\mathbf{s},\mathbf{a})(t)\delta(\mathbf{s},\mathbf{a}) = \left[-\frac{1}{\tau}\pi(\mathbf{s},\mathbf{a})\delta(\mathbf{s},\mathbf{a})\vartheta_{\mathbf{s},\mathbf{a}}(p_o)\right]\cdot\delta(\mathbf{s},\mathbf{a}) = -\frac{1}{\tau}\pi(\mathbf{s},\mathbf{a})\left[\delta(\mathbf{s},\mathbf{a})\right]^2\vartheta_{\mathbf{s},\mathbf{a}}(p_o) \leq 0.
$$

The equality only holds at the equilibrium point $\dot{p}_o(\mathbf{s},\mathbf{a})=0$. By Lemma 3, $\nabla V^{\pi}(\mathbf{s})\dot{\pi}(\mathbf{s},\mathbf{a})(t)\leq 0$. Therefore, according to Definition 2, the claim follows. ■

Theorem 4. Regardless of any initial value chosen for $p_n^0(\mathbf{s},\mathbf{a})$, and transferred knowledge $p_e(\mathbf{s},\mathbf{a})$, if the learning rate $\alpha(n)$, $\beta(n)$ and the transfer rate $\zeta(n)$ meets the required conditions meanwhile p_t and τ are sufficiently large, the Algorithm 1 converges.

Proof: The proof is the direct application of Theorem 2, which establishes the convergence given two conditions. First, the policy $p_o(\mathbf{s},\mathbf{a})$ tracks the solution of an ODE, by Theorem 1. Second, the tracked ODE has a strict Lyapunov function, by Theorem 3. Therefore, the learning process in Algorithm 1 converges. ■

V. NUMERICAL ANALYSIS

We validate the energy efficiency improvement of our proposed scheme by extensive simulations under practical configurations. Here, we simulate for a region consisting of three macro BSs and three micro BSs in an area of $1.5km\times 1.5km$ as Fig. 5 shows. Moreover, we assume that file transmission requests at location $x\in\mathcal{L}$ follow a Poisson point process with arrival rate $\lambda(x)$ and file size $1/\mu(x)=100$ kbyte. Beyond that, we assume the maximum transmission powers for BSs, i.e., 20W and 1W for macro and micro BSs, respectively. Based on the linear power consumption relationship in [6], the maximum operational powers for macro BS and micro BS are 865W and 38W, respectively. We set other main parameters in the propagation model according to the COST-231 modified Hata model [30] as summarized in Table I.

As for the proposed TACT algorithm, it's implemented with a discount factor $\gamma=0.001$ and the temperature $\tau=500$. Based on [14], the learning rate $\alpha(n)=1/n$ while $\beta(n)=1\Big/(n\log n)$. Moreover, the transfer rate $\zeta(n)=\theta^n$, with the transfer rate factor $\theta\in(0,1)$, thus satisfying the assumption in Theorem 1.

By the way, we define *cumulative energy consumption ratio* as the metric to test how much energy saving can be achieved due to the application of our proposed scheme. Specifically, we define the cumulative energy consumption ratio as: the ratio between the accumulative energy consumptions when certain BSs are turned off (as our scheme runs) and when all the BSs stay active since our simulation starts. Our definition is reasonable since it can show the foresighted energy efficiency improvement, which is exactly the goal of an energy saving scheme. Besides,

TABLE I

USED SIMULATION PARAMETERS

Parameter description		**Value**
Simulation area		1.5km $\times$ 1.5km
Maximum transmission power	Macro BS	20W
	Micro BS	1W
Maximum operational power	Macro BS	865W
	Micro BS	38W
Height	Macro BS	32m
	Micro BS	12.5m
Channel bandwidth		1.25MHz
Intra-cell interference factor		0.01
File requests	Arrival rate	$5 \times 10^{-6} \sim 10^{-4}$
	File size	100kbyte
Constant power percentage		0.1 $\sim$ 0.9

[a] For simplicity, we don't consider fast fading effect and noise influence in our simulation.

we compare the performance of our proposed schemes, i.e., classical actor-critic (AC) based and transfer actor-critic algorithm (TACT) based energy saving scheme with that of the state-of-the-art (SOTA) scheme, which assumes the controller can obtain a full knowledge of traffic loads in prior and find the optimal BS switching solution by exhausting all the possible ones.

A. Effect of traffic loads with static arrival rates on energy saving scheme

We firstly examine how much energy saving can be achieved versus different static traffic load arrival rates. [6] shows a homogeneous traffic distribution of $\lambda(x) = 10^{-4}$ for all $x \in \mathcal{L}$, which offers load corresponding to about 10% of BSs utilizations when all BSs are turned on. Therefore, we vary the homogeneous traffic arrival rate $\lambda(x)$ from 5×10^{-6} to 10^{-4}. Meanwhile, to compute the traffic load boundary points Γ_b, we record the average of traffic loads, i.e., Γ_a, in the whole region and then compute Γ_b for macro BSs and micro BSs by $\Gamma_{b,macro} = \frac{\Gamma_a}{3}$, $\Gamma_{b,micro} = \frac{1}{10}\Gamma_{b,macro}$, respectively.

Fig. 6 shows the effect of traffic load on energy savings when the portion of fixed power consumption q_i equals 0.5. Meanwhile, the transferred policy is generated from a source task

with the static arrival rate $\lambda = 5 \times 10^{-6}$. With the decrease of traffic load arrival rate λ from 10^{-4} to 5×10^{-6}, we can expect more significant energy conservation since if all the BSs stay active under lower traffic loads, the BSs are highly under-utilized. Moreover, the cumulative energy consumption ratio continues decreasing as the simulation runs since the controller will have a better understanding of the traffic loads, thereby knowing whichever action has a better energy efficiency. Unfortunately, since the proposed learning schemes are performed without the knowledge of traffic loads a prior, the performance of them are inferior to that of the SOTA scheme, especially at the beginning of the simulations. However, we can see that the gap compensated for the absent knowledge becomes smaller, when the TACT scheme is applied with the learned knowledge.

Fig. 7 presents the performance improvement[4] of TACT scheme over classical AC scheme. As expected, the TACT scheme yields a relatively large performance improvement, especially at the beginning of each simulation. In other words, the TACT scheme contributes to a performance jumpstart, or a faster convergence speed. Fig. 7 also depicts the similarity between the source task and the target task, measured by Kullback-Leibler divergence [31]. It shows a smaller Kullback-Leibler divergence between the source task and the target task leads to a more efficient transfer effect. Besides, we also plot the impact of transfer rate factor θ in Fig. 8. Generally speaking, as we expect, larger θ results in higher convergence rate and lower energy consumption ratio.

B. Effect of energy consumption models of BSs on energy saving scheme

In this part, we vary the portion of fixed power consumption q_i between 0 and 1, in order to cover various types of BSs with different energy consumption models. Fig. 9 shows the effect of energy consumption models of BSs on energy saving schemes when the traffic file request follows a homogeneous distribution with arrival rate $\lambda(x)$ from 5×10^{-6} to 10^{-4}. In this case, the transferred policy for a target task with a specific arrival rate λ is the learning result from a source task with the same arrival rate λ and an energy consumption model $q_i = 0.5$. As Fig. 9 depicts, the schemes will perform better when the constant power consumption accounts for a larger proportion of the whole energy consumption. The reason lies in that when the constant

[4]The performance improvement is calculated by dividing the energy consumption margin between TACT scheme and classical AC scheme over the energy consumption using classical AC scheme.

power consumption takes a larger percentage, i.e., $q_i = 0.9$, turning off one under-utilized BS will make a clearer difference and save more energy. On the other hand, more than half of the overall energy consumption usually takes place on the constant power, i.e., cooling, idle-mode signaling and processing in the present RAN infrastructure [7]. Therefore, our proposed scheme can render a strong positive effect in saving energy. It can also be found in Fig. 9, the performance of TACT scheme obviously outperforms that of the classical AC scheme in all the energy consumption configurations.

C. Performance of learning framework-based energy saving scheme in periodic traffic load scenario

In this section, we investigate the performance of the proposed scheme when traffic loads periodically fluctuates. [12] shows practical traffic load profile is periodical and can be approximated by a sinusoidal function $\overline{\lambda}(t) = \lambda_V \cdot \cos(2\pi(t+\phi)/D) + \lambda_M$, where t is the index of time, D is the period of a traffic load profile, λ_V is the variance of traffic profile and λ_M is the mean arrival rate. Therefore, we employ $\overline{\lambda}(t,x) = (0.99 \cdot \cos(2\pi(t+10)/24) + 1) \times 10^{-4}$ to approximate the practical traffic load arrival rate at location $x \in \mathcal{L}$. Fig. 10 compares the performance of the proposed schemes and shows that the TACT scheme converges faster than the classical AC scheme.

VI. Conclusion

In this paper, we have developed a learning framework for BS energy saving scheme. We specifically formulated the BS switching operation under a variant traffic load as a Markov decision process. Besides, we adopt the actor-critic method, a reinforcement learning approach to give the BS switching solution to decrease the overall energy consumption. Afterwards, to fully exploit the spatial and temporal relevancy in traffic loads, we propose a transfer actor-critic algorithm to improve the strategies by taking advantage of learned knowledge from neighboring regions or historical periods. Our proposed algorithm provably converges given certain restrictions that arise during the learning process, and the extensive simulation results manifest the effectiveness and robustness of our energy saving schemes under various practical configurations.

ACKNOWLEDGMENT

This paper is partially supported by the National Basic Research Program of China (973Green, Program No. 2012CB316000) and the National Natural Science Foundation of China (NSFC) under grant number 61071130.

APPENDIX

Proof of Theorem 1.

Proof: Without loss of generality, assume that at stage k, the state is $\mathbf{s}^k$ and the chosen action is $\mathbf{a}^k$. Moreover, the latest stage that the state-action pair $(\mathbf{s}^k, \mathbf{a}^k)$ occurred is stage m. Thus, by Algorithm 1, the policy $p_o^j(\mathbf{s}^k, \mathbf{a}^k)$ remain invariant for any $j \in [m, \cdots, k)$. For simplicity of representation, we denote one sequence $\hat{p}_o^{\hat{k}}(\mathbf{s}^k, \mathbf{a}^k) = p_o^k(\mathbf{s}^k, \mathbf{a}^k)$ and $\hat{p}_o^{\hat{k}-1}(\mathbf{s}^k, \mathbf{a}^k) = p_o^j(\mathbf{s}^k, \mathbf{a}^k)$, for any $j \in [m, \cdots, k)$, where the index $\hat{k}$ equals $\nu_2(\mathbf{s}^k, \mathbf{a}^k, k)$. In addition, the sequences $\hat{p}_n^{\hat{k}}(\mathbf{s}^k, \mathbf{a}^k)$ and $\hat{\delta}^{\hat{k}}(\mathbf{s}^k, \mathbf{a}^k)$ are defined analogously to $\hat{p}_o^{\hat{k}}(\mathbf{s}^k, \mathbf{a}^k)$. Thus, we have

$$\begin{aligned}
&\hat{p}_o^{\hat{k}}(\mathbf{s}^k, \mathbf{a}^k) = p_o^k(\mathbf{s}^k, \mathbf{a}^k) \\
&= \left[(1 - \zeta(\nu_2(\mathbf{s}^k, \mathbf{a}^k, k) - 1))p_n^k(\mathbf{s}^k, \mathbf{a}^k) + \zeta(\nu_2(\mathbf{s}^k, \mathbf{a}^k, k) - 1)p_e(\mathbf{s}^k, \mathbf{a}^k)\right]_{-p_t}^{p_t} \\
&= \left[(1 - \zeta(\hat{k} - 1))\hat{p}_n^{\hat{k}}(\mathbf{s}^k, \mathbf{a}^k) + \zeta(\hat{k} - 1)p_e(\mathbf{s}^k, \mathbf{a}^k)\right]_{-p_t}^{p_t}.
\end{aligned} \tag{16}$$

Firstly, assume that p_t is large enough such that $\left|p_o^k(\mathbf{s}^k, \mathbf{a}^k)\right| < p_t$ and $\left|p_o^{k+1}(\mathbf{s}^k, \mathbf{a}^k)\right| < p_t$, while the assumption will be dropped later.

Subtracting Equation (10) to Equation (16), we obtain

$$\begin{aligned}
&\hat{p}_o^{\hat{k}+1}(\mathbf{s}^k, \mathbf{a}^k) - \hat{p}_o^{\hat{k}}(\mathbf{s}^k, \mathbf{a}^k) \\
&= (1 - \zeta(\hat{k} - 1)) \left(\hat{p}_n^{\hat{k}+1}(\mathbf{s}^k, \mathbf{a}^k) - \hat{p}_n^{\hat{k}}(\mathbf{s}^k, \mathbf{a}^k)\right) - (\zeta(\hat{k}) - \zeta(\hat{k} - 1)) \left(\hat{p}_n^{\hat{k}+1}(\mathbf{s}^k, \mathbf{a}^k) - p_e(\mathbf{s}^k, \mathbf{a}^k)\right) \\
&= -\beta(\hat{k})(1 - \zeta(\hat{k} - 1))\hat{\delta}^{\hat{k}}(\mathbf{s}^k, \mathbf{a}^k) - (\zeta(\hat{k}) - \zeta(\hat{k} - 1)) \left(\hat{p}_n^{\hat{k}+1}(\mathbf{s}^k, \mathbf{a}^k) - p_e(\mathbf{s}^k, \mathbf{a}^k)\right).
\end{aligned} \tag{17}$$

The last equality holds because of Equation (9).

Define $t_0 = 0$ and $t_{\hat{k}} = \sum_{j=0}^{\hat{k}-1} \beta(j)$. For $t \geq 0$, let $\mathfrak{K}(t)$ denote the unique value of $\hat{k}$ such that $t_{\hat{k}} \leq t < t_{\hat{k}+1}$, as Fig. 4-(a) depicts. For $t < 0$, set $\mathfrak{K}(t) = 0$. Define the *continuous time interpolation* $\hat{p}_{\mathbf{s}^k, \mathbf{a}^k}^0(\cdot)$ on $(-\infty, \infty)$ by $\hat{p}_{\mathbf{s}^k, \mathbf{a}^k}^0(t) = p_o^0(\mathbf{s}^k, \mathbf{a}^k)$ for $t \leq 0$, and for $t \geq 0$,

$$\hat{p}_{\mathbf{s}^k, \mathbf{a}^k}^0(t) = \hat{p}_o^{\hat{k}}(\mathbf{s}^k, \mathbf{a}^k), \text{ for } t_{\hat{k}} \leq t < t_{\hat{k}+1}.$$

Moreover, we define the *sequence of shifted processes* $\hat{p}^{\hat{k}}_{\mathbf{s}^k,\mathbf{a}^k}(t) = \hat{p}^{0}_{\mathbf{s}^k,\mathbf{a}^k}(t_{\hat{k}}+t),\ t\in(-\infty,\infty)$, as Fig. 4-(d) depicts. Define $Y_j = 0$ and $Z_j = 0$ for $j < 1$. Moreover, define $Y_j = (1-\zeta(j-1))\hat{\delta}^j(\mathbf{s}^k,\mathbf{a}^k)$ and $Z_j = (\zeta(j)-\zeta(j-1))\left(\hat{p}^{j+1}_n(\mathbf{s}^k,\mathbf{a}^k) - p_e(\mathbf{s}^k,\mathbf{a}^k)\right)$ for $j \geq 1$. Define $Z^0(t) = 0$ for $t \leq 0$ and

$$Z^0(t) = \sum_{j=0}^{\mathfrak{K}(t)-1} Z_j,\ Z^{\hat{k}}(t) = Z^0(t_{\hat{k}}+t) - Z^0(t_{\hat{k}}) = \sum_{j=\hat{k}}^{\mathfrak{K}(t_{\hat{k}}+t)-1} Z_j,\ t \geq 0.$$

Taking into count the definitions above (recall that $\mathfrak{K}(t_{\hat{k}}) = \hat{k}$), the following equation can be achieved by a manipulation of Equation (17)

$$\hat{p}^{\hat{k}}_{\mathbf{s}^k,\mathbf{a}^k}(t) = \hat{p}^{\hat{k}}_o(\mathbf{s}^k,\mathbf{a}^k) - \sum_{j=\hat{k}}^{\mathfrak{K}(t_{\hat{k}}+t)-1} (\beta(j)Y_j + Z_j) = \hat{p}^{\hat{k}}_o(\mathbf{s}^k,\mathbf{a}^k) - \sum_{j=\hat{k}}^{\mathfrak{K}(t_{\hat{k}}+t)-1} (\beta(j)Y_j) - Z^{\hat{k}}(t). \quad (18)$$

Since $\hat{p}^{\hat{k}}_{\mathbf{s}^k,\mathbf{a}^k}(t)$ is piecewise constant, we can rewrite Equation (18) as

$$\hat{p}^{\hat{k}}_{\mathbf{s}^k,\mathbf{a}^k}(t) = \hat{p}^{\hat{k}}_o(\mathbf{s}^k,\mathbf{a}^k) - \int_0^t Y_{\mathfrak{K}(t_{\hat{k}}+x)}dx - Z^{\hat{k}}(t) + \varphi^{\hat{k}}(t), \quad (19)$$

where $\varphi^{\hat{k}}(t)$ is the outcome due to the replacement of the first sum in Equation (18) by an integral. $\varphi^{\hat{k}}(t) = 0$ at the times when the interpolated sequences have jumps, i.e., $t = t_{\hat{k}'} - t_{\hat{k}},\ \hat{k}' > \hat{k}$, and $\varphi^{\hat{k}}(t) \to 0$ in t as $\hat{k} \to \infty$ under the assumption in Equation (13).

Besides that, by our assumption that $\lim \zeta(\hat{k})/\beta(\hat{k}) \to 0$ as $\hat{k} \to \infty$, $Z_{\hat{k}} = (\zeta(\hat{k}) - \zeta(\hat{k}-1)) \cdot \left(\hat{p}^{\hat{k}+1}_n(\mathbf{s}^k,\mathbf{a}^k) - p_e(\mathbf{s}^k,\mathbf{a}^k)\right) = o(\beta(\hat{k}))\left(\hat{p}^{\hat{k}+1}_n(\mathbf{s}^k,\mathbf{a}^k) - p_e(\mathbf{s}^k,\mathbf{a}^k)\right)$. Therefore, $Z^{\hat{k}}(t) = \sum_{j=\hat{k}}^{\mathfrak{K}(t_{\hat{k}}+t)-1} o(\beta(j))\left(\hat{p}^{j+1}_n(\mathbf{s}^k,\mathbf{a}^k) - p_e(\mathbf{s}^k,\mathbf{a}^k)\right)$. Thus, as $\hat{k} \to \infty$, $Z^{\hat{k}}(t)$ is negligible, since it's a small order of magnitude to $\sum_{j=\hat{k}}^{\mathfrak{K}(t_{\hat{k}}+t)-1} \beta(j)Y_j$.

Given the above discussion, as $\hat{k} \to \infty$, the sequence of functions $\hat{p}^{\hat{k}}_{\mathbf{s}^k,\mathbf{a}^k}(t) = \hat{p}^{\hat{k}}_o(\mathbf{s}^k,\mathbf{a}^k) - \int_0^t Y_{\mathfrak{K}(t_{\hat{k}}+x)}dx$ is equicontinous. Hence, by the Arzelà-Ascoli Theorem [29], there is a convergent subsequence in the sense of uniform convergence on each bounded time integral, and it's easily seen that any limit of $\hat{p}_{\mathbf{s}^k,\mathbf{a}^k}(t)$, or the discrete equivalent $\hat{p}^{\hat{k}}_o(\mathbf{s}^k,\mathbf{a}^k)$, must track the solution of the ODE $\dot{\hat{p}}_{\mathbf{s}^k,\mathbf{a}^k}(t) = -\hat{\delta}^{\hat{k}}(\mathbf{s}^k,\mathbf{a}^k)$ for sufficiently large $\hat{k}$.

Next, in the special case where $\hat{p}^{\hat{k}-1}_o(\mathbf{s}^k,\mathbf{a}^k) = p_t$ and $\hat{\delta}^{\hat{k}-1}(\mathbf{s}^k,\mathbf{a}^k) \geq 0$, at next stage k, the overall policy $\hat{p}^{\hat{k}}_o(\mathbf{s}^k,\mathbf{a}^k)$ would equal p_t. Thus, the ODE $\dot{\hat{p}}_{\mathbf{s}^k,\mathbf{a}^k}(t) = 0$. Similar discussion can be easily applied to the case, where $\hat{p}^{\hat{k}-1}_o(\mathbf{s}^k,\mathbf{a}^k) = -p_t$ and $\hat{\delta}^{\hat{k}-1}(\mathbf{s}^k,\mathbf{a}^k) \leq 0$.

Furthermore, as $k \to 0$, by Lemma 1, $\hat{k} = \nu_2(\mathbf{s},\mathbf{a},k) \to \infty$.

Summarizing the above discussion and taking into account $\delta(\mathbf{s}^k, \mathbf{a}^k) = \lim \delta^k(\mathbf{s}^k, \mathbf{a}^k)$ as $k \to \infty$, we can obtain

$$\dot{p}_o(\mathbf{s}^k, \mathbf{a}^k)(t) = -\delta(\mathbf{s}^k, \mathbf{a}^k)\vartheta_{\mathbf{s}^k, \mathbf{a}^k}(p_o). \tag{20}$$

The claim follows. ■

REFERENCES

[1] J. Wu, S. Rangan, and H. Zhang, *Green Communications - Theoretical Fundamentals, Algorithms and Applications*. CRC Press, Sep. 2012.

[2] H. Zhang, A. Gladisch, M. Pickavet, Z. Tao, and W. Mohr, "Energy efficiency in communications," *IEEE Communications Magazine*, vol. 48, no. 11, pp. 48–49, Nov. 2010.

[3] M. Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo, "Optimal energy savings in cellular networks," in *Proceedings of IEEE ICC Workshops 2009*, Dresden, Germany, Jun. 2009, pp. 1–5.

[4] China Mobile Research Institute, "C-RAN: road towards green radio access network," Tech. Rep., 2010.

[5] G. P. Fettweis and E. Zimmermann, "ICT energy consumption-trends and challenges," in *Proceedings of WPMC 2008*, vol. 4, Lapland, Finland, Sep. 2008, p. 6.

[6] K. Son, H. Kim, Y. Yi, and B. Krishnamachari, "Base station operation and user association mechanisms for energy-delay tradeoffs in green cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 8, pp. 1525 –1536, Sep. 2011.

[7] C. Peng, S.-B. Lee, S. Lu, H. Luo, and H. Li, "Traffic-driven power savings in operational 3G cellular networks," in *Proceedings of ACM Mobicom 2011*, Las Vegas, Nevada, USA, Sep. 2011, p. 121132.

[8] Z. Niu, "TANGO: traffic-aware network planning and green operation," *IEEE Wireless Communications*, vol. 18, no. 5, pp. 25 –29, Oct. 2011.

[9] L. Chiaraviglio, D. Ciullo, M. Meo, M. Marsan, and I. Torino, "Energy-aware UMTS access networks," in *Proceedings of WPMC 2008*, Lapland, Finland, Sep. 2008.

[10] Z. Niu, Y. Wu, J. Gong, and Z. Yang, "Cell zooming for cost-efficient green cellular networks," *IEEE Communication Magazine*, vol. 48, no. 11, pp. 74–79, Nov. 2010.

[11] R. Li, Z. Zhao, Y. Wei, X. Zhou, and H. Zhang, "GM-PAB: a grid-based energy saving scheme with predicted traffic load guidance for cellular networks," in *Proceedings of IEEE ICC 2012*, Ottawa, Canada, Jun. 2012, p. 11601164.

[12] E. Oh and B.Krishnamachari, "Energy savings through dynamic base station switching in cellular wireless access networks," in *Proceedings of IEEE Globecom 2010*, Miami, Florida, USA, Dec. 2010, p. 15.

[13] S. Zhou, J. Gong, Z. Yang, Z. Niu, and P. Yang, "Green mobile access network with dynamic base station energy saving," in *Proceedings of ACM Mobicom 2009*, Beijing, China, Sep. 2009.

[14] V. Konda and V. Borkar, "Actor-criticstype learning algorithms for markov decision processes," *SIAM Journal on Control and Optimization*, vol. 38, no. 1, pp. 94–123, 1999.

[15] V. Konda and J. Tsitsiklis, "Actor-critic algorithms," *SIAM Journal on Control and Optimization*, vol. 42, no. 4, pp. 1143–1166, 2000.

[16] R. Sutton and A. Barto, *Reinforcement learning: An introduction*. Cambridge Univ Press, 1998.

[17] H. Berenji and D. Vengerov, "A convergent actor-critic-based FRL algorithm with application to power management of wireless transmitters," *IEEE Transactions on Fuzzy Systems*, vol. 11, no. 4, pp. 478–485, 2003.

[18] F. Woergoetter and B. Porr, "Reinforcement learning," *Scholarpedia*, vol. 3, no. 3, p. 1448, 2008.

[19] S. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[20] D. Aha, M. Molineaux, and G. Sukthankar, "Case-based reasoning in transfer learning," *Case-Based Reasoning Research and Development*, pp. 29–44, 2009.

[21] J. Celiberto, Luiz A., J. P. Matsuura, R. L. de Mantaras, and R. A. C. Bianchi, "Using transfer learning to speed-up reinforcement learning: a cased-based approach," in *Proceedings of the 2010 Latin American Robotics Symposium and Intelligent Robotics Meeting*, Washington, DC, USA, Oct. 2010.

[22] M. Taylor and P. Stone, "Transfer learning for reinforcement learning domains: A survey," *The Journal of Machine Learning Research*, vol. 10, pp. 1633–1685, Jul. 2009.

[23] X. Zhou, Z. Zhao, R. Li, Y. Zhou, and H. Zhang, "The predictability of cellular networks traffic," in *Proceedings of IEEE ISCIT 2012*, Gold Coast, Australia, Oct. 2012.

[24] R. Li, Z. Zhao, X. Chen, and H. Zhang, "Energy saving through a learning framework in greener cellular radio access networks," in *Proceedings of IEEE Globecom 2012*, Anaheim, USA, Dec. 2012.

[25] A. Sang, X. Wang, M. Madihian, and R. Gitlin, "Coordinated load balancing, handoff/cell-site selection, and scheduling in multi-cell packet data systems," *Wireless Networks*, vol. 14, no. 1, pp. 103–120, 2008.

[26] H. Kim, G. De Veciana, X. Yang, and M. Venkatachalam, "Alpha-optimal user association and cell load balancing in wireless networks," in *Proceedings of IEEE INFOCOM 2010*, San Diego, CA, USA, Mar. 2010, pp. 1–5.

[27] K. Zhou, "Robust cross-layer design with reinforcement learning for IEEE 802.11n link adaptation," in *Proceedings of IEEE ICC 2011*, Kyoto, Japan, Jun. 2011.

[28] S. Singh, T. Jaakkola, M. Littman, and C. Szepesvri, "Convergence results for single-step on-policy reinforcement-learning algorithms," *Machine Learning*, vol. 38, no. 3, pp. 287–308, 2000.

[29] H. Kushner and G. Yin, *Stochastic approximation and recursive algorithms and applications*, 2nd ed. New York, USA: Springer, 2003, vol. 35.

[30] IEEE 802.16 Boradband Wireless Access Working Group, "IEEE 802.16m evaluation methodology document (EMD)," Jul. 2008. [Online]. Available: http://ieee802.org/16

[31] S. Kullback and R. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

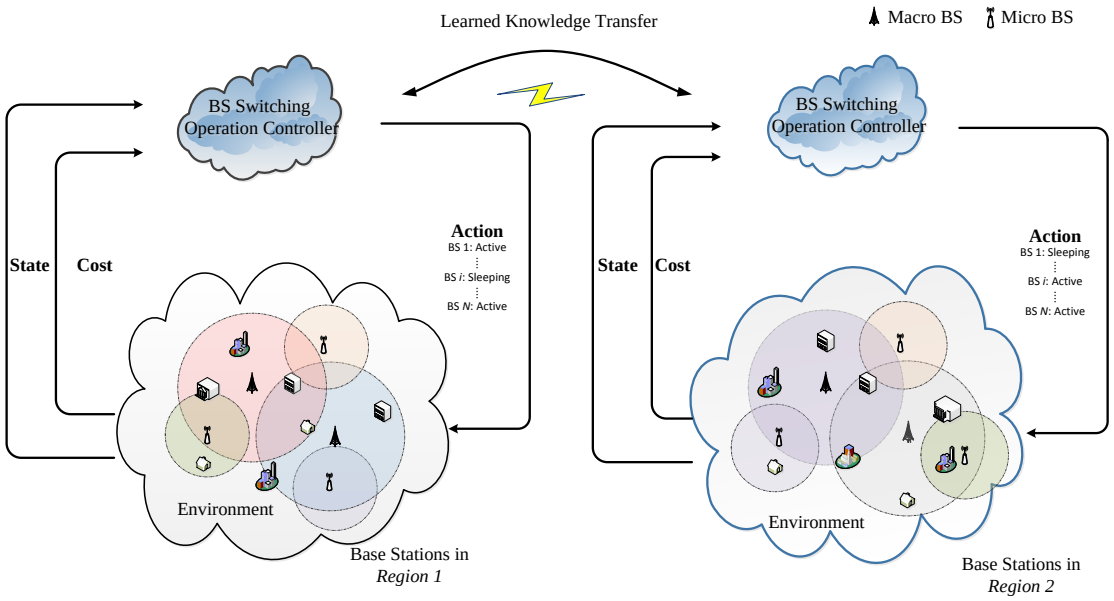

Fig. 1. Transfer learning for reinforcement learning in BS switching operation scenario.

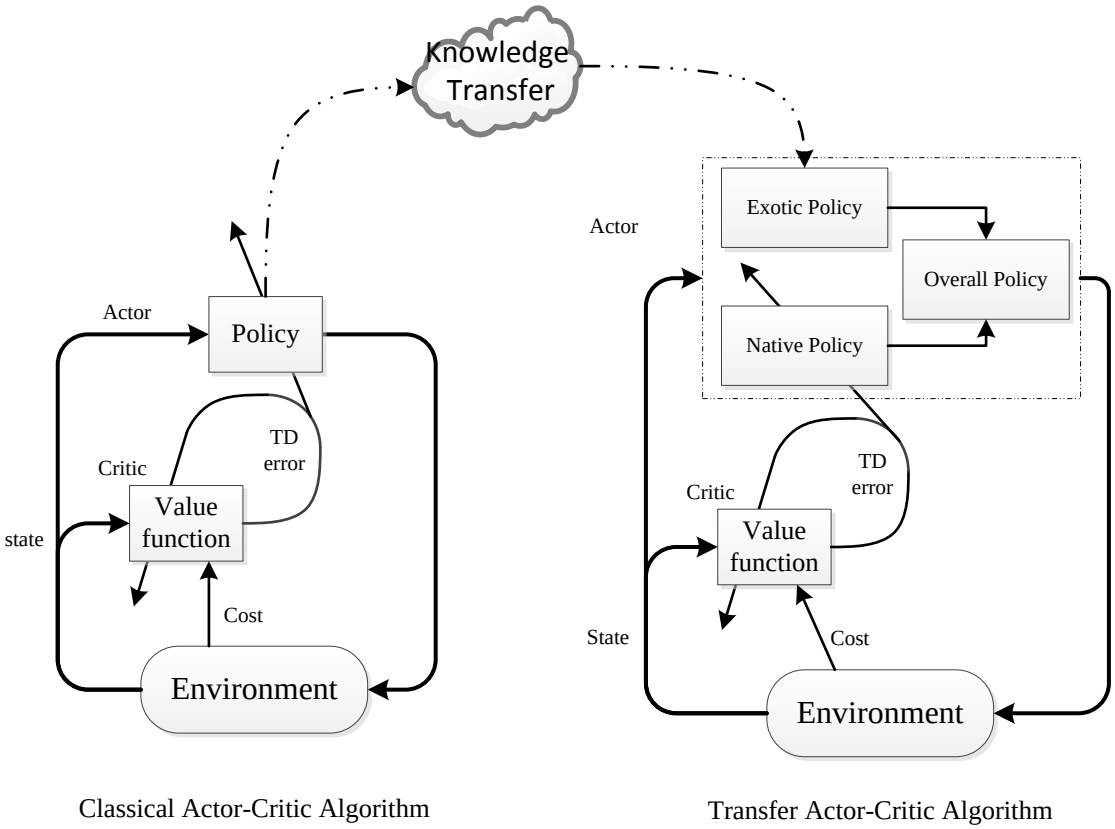

Fig. 2. Architecture of classical actor-critic algorithm and transfer actor-critic algorithm (TACT).

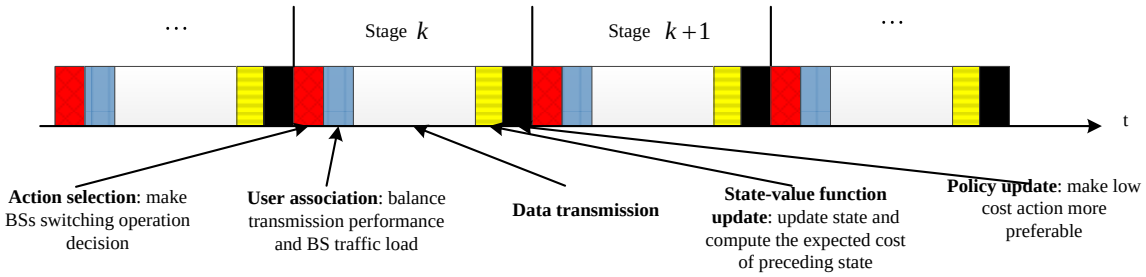

Fig. 3. Illustration of actor-critic learning framework for energy saving scheme.

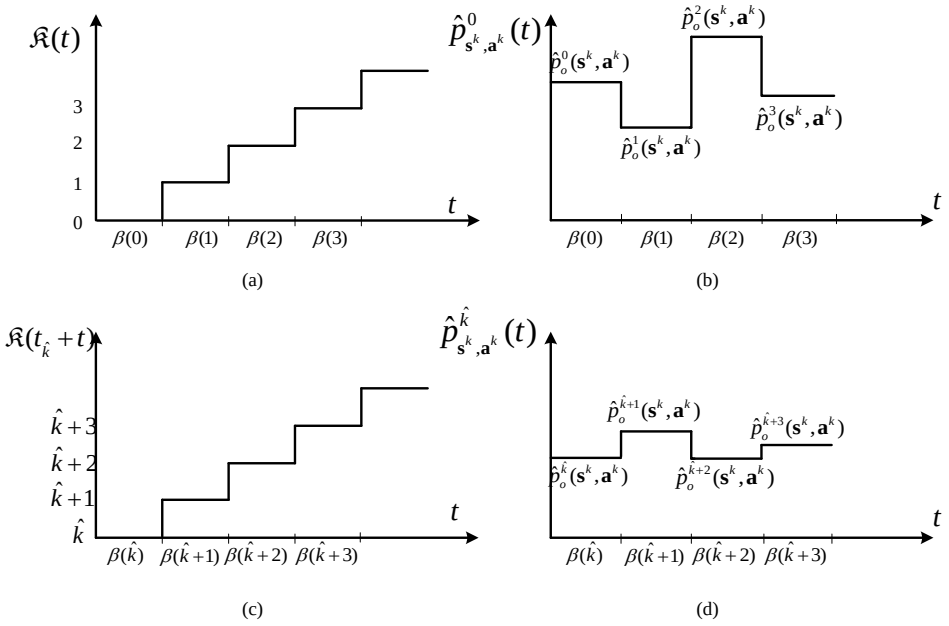

Fig. 4. Illustration of (a) the function $\Re(t)$, (b) the function $\hat{p}^0_{\mathbf{s}^k,\mathbf{a}^k}(t)$, (c) the function $\Re(t_{\hat{k}}+t)$ and (d) the function $\hat{p}^{\hat{k}}_{\mathbf{s}^k,\mathbf{a}^k}(t)$.

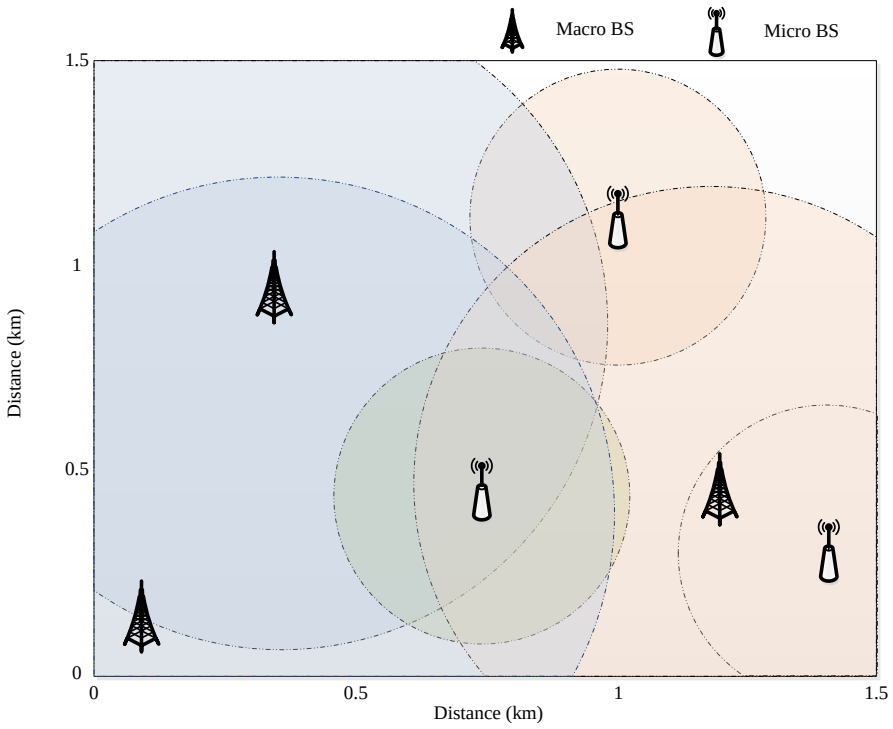

Fig. 5. Illustration of BS deployment in our simulation scenario.

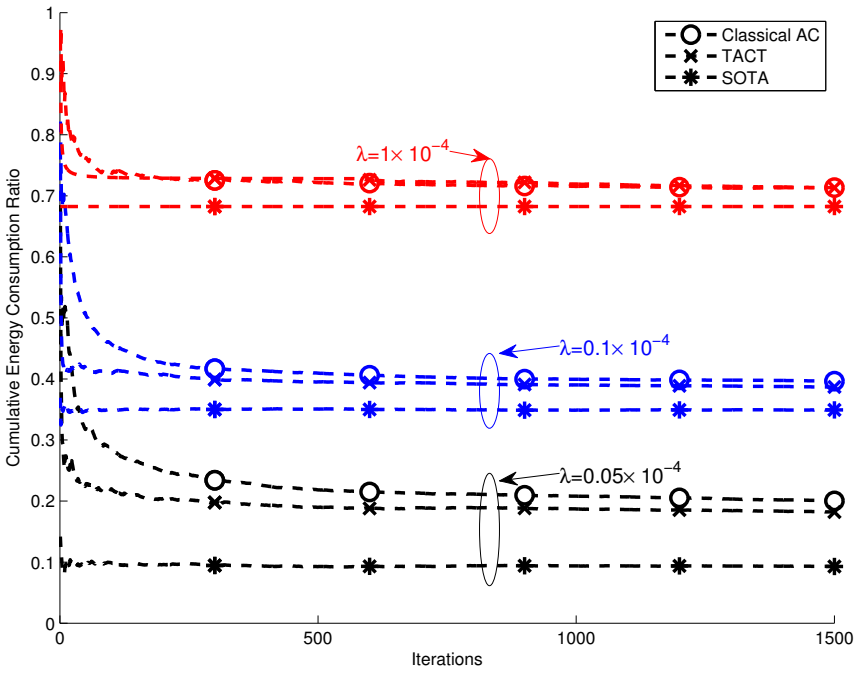

Fig. 6. Performance comparison among classical AC scheme, TACT scheme and SOTA scheme under various homogeneous traffic arrival rates when the transfer rate $\theta = 0.1$.

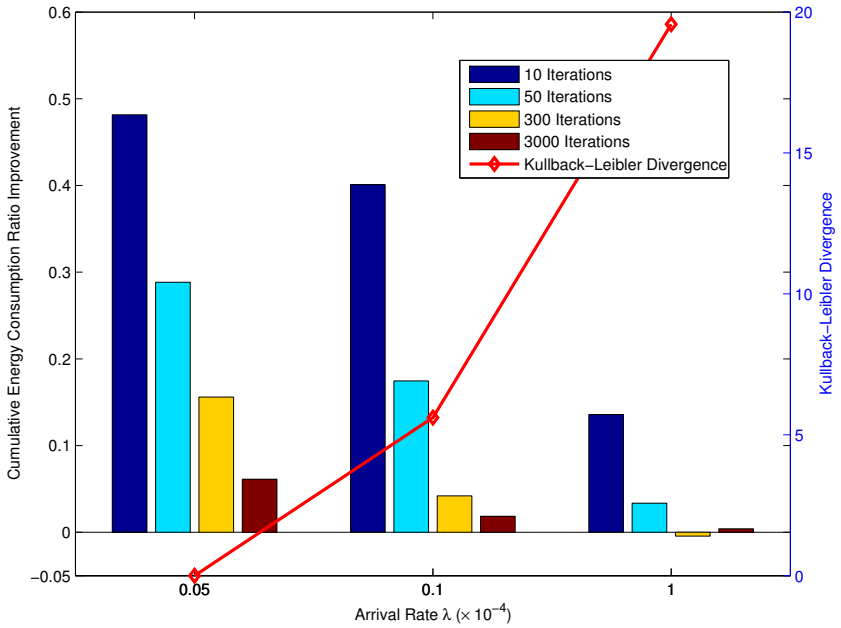

Fig. 7. Performance improvement of TACT scheme over classical AC scheme versus Kullback-Leibler divergence when the transfer rate $\theta = 0.1$.

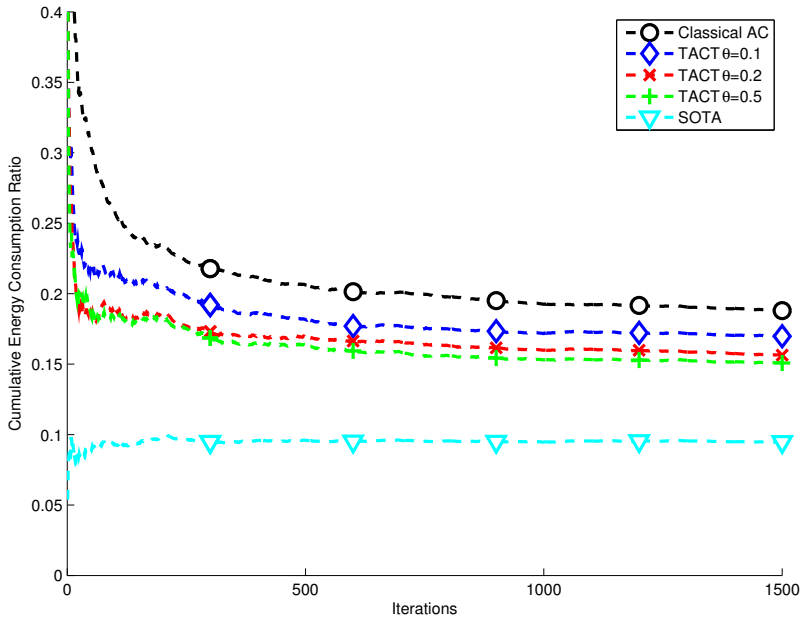

Fig. 8. Performance impact of the transfer rate factor θ to the TACT scheme when $\lambda = 5 \times 10^{-6}$.

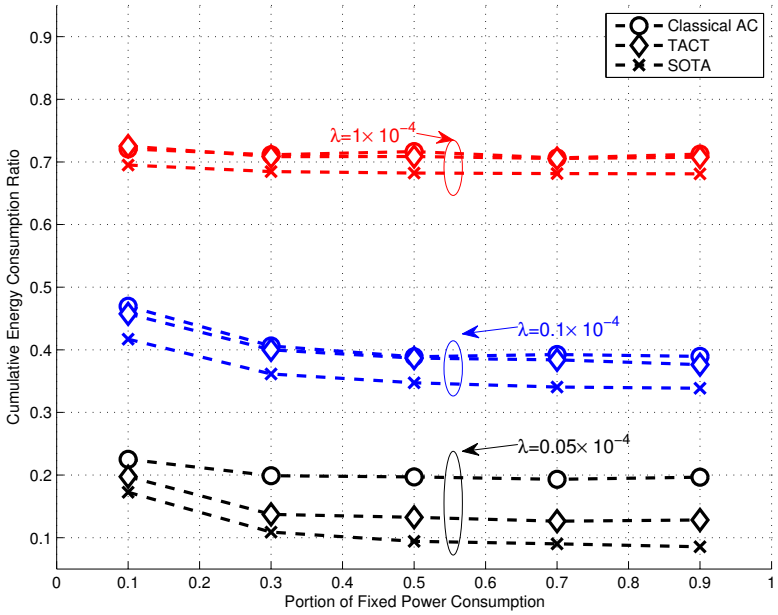

Fig. 9. Performance comparison among classical AC scheme, TACT scheme, SOTA scheme under different energy consumption models after 1500 iterations when the transfer rate $\theta = 0.1$.

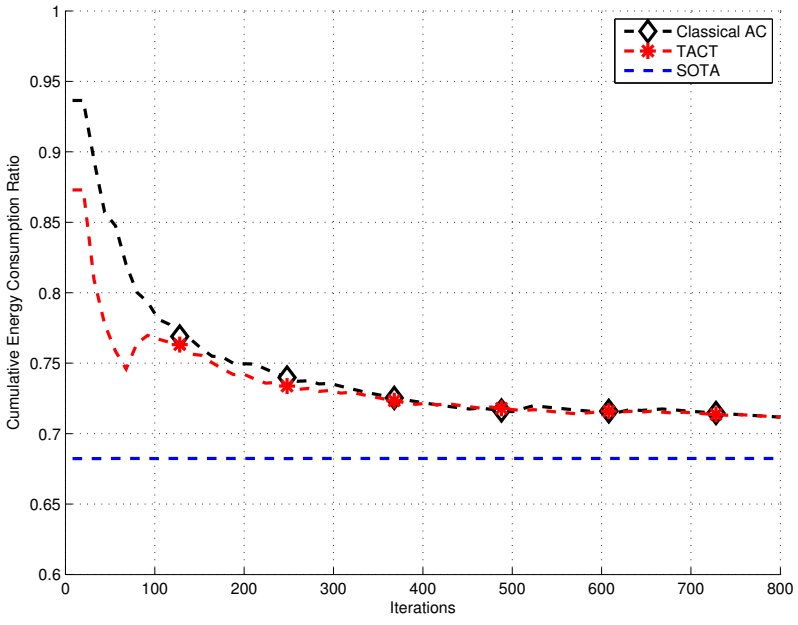

Fig. 10. Performance comparison of classical AC scheme, TACT scheme, SOTA scheme with time-variant traffic arrival rate $\overline{\lambda}(t,x) = (0.99 \cdot \cos(2\pi(t+10)/24) + 1) \times 10^{-4}$ when $\theta = 0.1$.